C h a p t e r Eleven

A STATISTICAL ANALYSIS OF LEXIS IN CONVERSATIONAL ENGLISH*

MATERIAL

The material for the statistical analysis of lexis to be presented is taken from the transcriptions of English conversations by Svartvik and Quirk [1980, the London-Lund Corpus], prosodically analysed by the editors. The transcripts comprise 51 surrepetitious texts of about 175,000 running words recorded in the period 1953-1976. The recordings represent specimens of spontaneous conversation among British speakers, aged 18-62, educated to the university level. The analysis is based on ten samples of conversational English.

HYPOTHESIS

When I first read the texts, some of them struck me as vivid, varied, linguistically interesting, while others sounded formulaic and flat. In my 1987 book I proposed that this fact can be accounted for in terms of the concept of *linguistic activeness* as shown by a speaker in an interactional context. It is proposed here that the linguistic activeness of a speaker can be evaluated in terms of three sets of criteria: (a) communicative criteria connected with the type of interaction, participants, setting, topic, etc., (b) discourse criteria referring to the flow

---

* Barbara Lewandowska-Tomaszczyk, University of Łódź.

of the conversation (number and type of turns, verbosity of speakers, etc.), as well as (c) narrow linguistic criteria, expressed, among other things, by some quantitative characteristics of the lexis the conversationalists use in their language. The linguistic criteria of activeness also cover less conventional constructions and uses of words and phrases.

The analysis is based on ten samples of conversational English, five of which (group I, Table 1) represent the English of five different speakers who are intuitively evaluated as 'good' and interesting conversationalists, while the other five were taken from speakers who are judged poor and dull conversationalists (group II, Table 1).

Table 1

GI 'good'					GII 'poor'				
	LO_1	LD_1	LV_1	LS_1		LO_2	LD_2	LV_2	LS_2
1)	21.0	45.5	33.1	0.9	6)	6.9	31.4	23.8	0.4
2)	16.9	44.5	32.1	0.6	7)	2.9	33.5	29.5	0.6
3)	15.7	44.7	27.8	0.8	8)	2.6	34.6	25.6	0.1
4)	15.7	42.4	26.8	0.4	9)	2.5	40.8	27.6	0.1
5)	9.7	38.2	27.1	0.4	10)	2.1	35.4	29.5	0.6

GI - data (%) for 'good' samples

GII - data (%) for 'poor' samples

METHOD

Each sample is 1500 words long and each covers the beginning of the interaction. The analysis is based on the evaluation of four types of variable (cf. Linnarud, 1983 for similar variables; note that what are termed LO and LS in the Linnarud study are understood here in a different way).

Linguistic Originality (LO)

LO covers the percentage of witticisms and instances of not fully conventionalised figurative language, such as:

(1) f e m a l e a c a d e m i c: I used to sew a lot in the days *when I was a human being*

(2) I *spoil* my ballot paper every time I vote by putting a TICK instead of a cross

(3) A: they were rather *parochial* down here, B: in an *international* way

(4) the best way to get a job is *not to care* whether you get it or not

(5) I had enough of his *elaborate porridge*

(6) (talking about some exotic food): I'm sure it was (mu/ mu/) -- cos it sounded like mush mush, you know -- as in *driving SLEIGH dogs across the frozen WASTES*

(7) n o n-a c a d e m i c f e m a l e: I mean I've got a thing anyway about academic women. I think something ghastly happens to them -- but these women -- you know -- *untouched by human hand,* it s just frightening.

The conversational contributions included in this group do not necessarily have to be aesthetically pleasing:

(8) (two medical students talking about a colleague sitting the internal diseases exam)
A: then he, then he started talking and talking
B: *verbal diarrhoea* in fact

Lexical Density (LD)

LD expresses the percentage of the total number of words in a text that are *lexical words* (nouns, verbs - excluding auxiliaries, adjectives, non-deictic adverbs). Proper nouns as well as the lexical verbs *be, have,* and *do* were also included:

(9) I *am* a teacher (2 lexical words)

(10) I am *standing* here (1 lexical word)

LD is not affected by the length of the text (e.g., if we divide a 1500 words sample into three 500 words portions, the percentage of lexical words in each portion is approximately stable, (i.e., around 110-120 words)).

Lexical Variation (LV)

LV expresses the type/token ratio of the lexis in the language of each speaker. LV is negatively correlated, as will be

seen in Table 4, with the length of the text. One of the reasons is that it is easier to avoid repeating words in a shorter than in a longer text. LV is also dependent on another parameter, i.e. the communicative modality: LV is much higher in *written* than in *spoken* texts (cf. Table 2).

T a b l e 2

Comparison of LD and LV ranges in spoken and written English

	LD	LV
Spoken English (conversations)	31.4 - 40.8	23.8 - 29.5
Written English (compositions) [after Linnarud 1983]	33.0 - 54.0	61.0 - 88.0

The authors tend not to repeat the same word in the written text - unless it is really necessary, while the spoken language, especially in spontaneous conversations like the ones under analysis, cannot be controlled to that extent.

Lexical Sophistication (LS)

LS covers the percentage of lexical words with the frequency counts 1 to 5 per 1 million or lower (per 4 million and per 18 million) as indicated in the Thorndike and Lorge (1968) frequency lists. Thorndike and Lorge provide frequencies based on *written* sources. There may be some justification in assuming that if a word is listed with such a low frequency for written data, it may be still less frequent in spoken material. On the other hand, however, some words occurring in conversational language may not be used in written varieties at all.

Here are some examples of words with low frequencies and those not listed in Thorndike and Lorge:

(11) *disseminate* (2/1 mil.), *menial* (2/1 mil.), *criterion* (1/1 mil.), *prologue* (5/1 mil.), *excruciating* (1/1 mil), *scepticism* (1/1 mil.), *conceivable* (5/1 mil.) *prerequisite* (9/4 mil.), *authoritarian* (4/18 mil.), *scrupulosity* (4/18 mil.)

not listed:

syrupiness

teachy (English is a very teachy subject)

studenty (they were sort of research studenty kind of people)

Lexical creativity then enters the class of LS.

STATISTICS

COMPARING MEANS (t-test for independent samples)

Group I (GI) in Table 1 above presents values for the four variables discussed in the preceding section: LO, LD, LV, and LS for five 'good' speakers. Group II gives values of the same variables for the 'poor' speakers.

To calculate averages (av), standard deviations (s), and t-values (t) for 4 and 8 degrees of freedom (Df), I have used the IFASTATS package of computer programs written by D. Coleman and described in this volume. A Commodore-64 microcomputer was used to conduct the calculations.

Differences in the means of LO and LD between GI and GII appear to be statistically significant at the 0.005% level ($p \leq 0.005$). Such a level of probability indicates that the differences in the averages are very highly significant.

LS differences, on the other hand, are significant only at the 10% level (approaching the 5% level), which is considered not too stringent a significance level.

LV is even more controversial. The t-value is not significant even at 10% level, though it is close to the critical value at that level (1.397). The null hypothesis in such a situation cannot be rejected for this set of observation, and the analysis will have to be repeated with a larger sample. On the other hand, such a result may be indicative of the fact that LV is a variable which differs significantly not in the texts of different creativity but which, as has been mentioned above, may reflect differences of other types, e.g. those between spoken and written language (cf. Table 2).

CORRELATION (Spearman's *rho*)

The comparison of all the results presented in Group I and

Table 3

IFA MEANS (t-test)

LO		
GI av 15.8 s 4.04 Df 4	GII av 3.4 s 1.97 Df 4	GI/GII s 2.03 t 6.157 Df 8 ($p \leq 0.005$)
LD		
GI av 43.0 s 2.94 Df 4	GII av 35.14 s 3.50 Df 4	GI/GII s 2.04 t 3.868 Df 8 ($p \leq 0.005$)
LV		
GI av 29.38 s 2.98 Df 4	GII av 27.2 s 2.49 Df 4	GI/GII s 1.73 t 1.253 Df 8 (-) critical value at $p \leq 0.10$ 1.397
LS		
GI av 0.62 s 0.22 Df 4	GII av 0.36 s 0.25 Df 4	GI/GII s 0.15 t 1.714 Df 8 ($p \leq 0.10$)

Group II seems to point to some kind of a systematic relationship between the values of LO, LD, LV, and LS for all the analysed samples. In order to find out whether there is a statistically significant correlation between these variables, i.e., whether, say, the higher values of LO go together with the higher values of LD or LV, or not, one of the simpler quantitative assessments of correlation has been used. It is the *Spearman's rho* correlation measure, based on ranking. IFASTATS was used here as well, but as will be seen below, the calculation is much simpler than in the case of, say, the t-test, so it can be easily done without a computer.

Table 4 gives us all the steps necessary to test the Sperman's rho correlation between LO and LD.

T a b l e 4

Calculation of Spearman's *rho* correlation coefficient

LO - LD			
r a n k s		d	d^2
1	1	0	0
2	3	-1	1
3.5	2	1.5	2.25
3.5	4	-0.5	0.25
5	5	-1	1
6	10	-4	16
7	9	-2	4
8	8	0	0
9	5	4	16
10	7	3	9
			$\Sigma = 49.5$

First the set of all 10 scores of LO are ranked from largest to smallest (or from smallest to largest). If there are two or more *tied scores*, i.e., scores of the same values (e.g. 3/ and 4/), the mean of the ranks is used (3.5 for each) that would have been occupied if no tie had occurred. Similarly we rank the scores on the other variable (LS). The difference (d) between the ranks is then calculated and squared (d^2) for each pair of scores. We calculate the sum of all the squares -- *sigma* (Σ = = 49.5), and by substituting the values in the formula below (5) we calculate the Spearman correlation coefficient.

To find out whether the correlation is statistically significant we check statistical tables for critical values for *rho* for a given number of pairs of observations (10 - in our case). Statistical tables will also give us information concerning the probability level for a given result. Table 5 below gives us rankings and values of the Spearman correlation coefficient *rho*

$$rho = 1 - \frac{6 \Sigma d^2}{N (N^2 - 1)}$$

d rank difference
Σ sum of squares
N number of pairs of observation

$$rho = 1 - \frac{6 \times 49,5}{10 (10^2 - 1)} = 0.700 \quad (p \leq 0.025 \text{ for } N = 10)$$

Fig. 1 Spearman s *rho* formula

between LO-LV, LO-LS, LD-LV, LD-LS, LV-LS, as well as between LV and the length of the text.

T a b l e 5

Correlations

LO - LV			LO - LS			LD - LV		
1	1		1	1		1	1	
2	2		2	4		3	2	
3.5	2		3.5	2		2	5	
3.5	8		3.5	7		4	8	
5	7		5	7		6	7	
6	10		6	7		10	10	
7	3.5		7	4		9	3.5	
8	9		8	9.5		8	9	
9	6		9	9.5		5	6	
10	3.5	rho = 0.352	10	4	rho = 0.570	7	3.5	rho = 0.494
(-)	crit.val.at p ≤ 0.10	0.455			(p ≤ 0.05)			(p ≤ 0.10)

LD - LS			LV - LS		LV - no of words
1	1		1	1	e.g. 3) 190 (45%) - I 500
3	4		2	4	130 (31%) - II 500
2	2		5	2	97 (24%) - III 500
4	7		8	7	
6	7		7	7	= - 0.914 (0.05)
10	7		10	7	
9	4		3.5	4	
8	9.5		9	9.5	
5	9.5		6	9.5	
7	4	rho = 0.537	3.5	4	rho = 0.781
		(p ≤ 0.10)			(p ≤ 0.025)

Results

The results confirm some correlations. The correlation between LO and LD is significant at the 2.5% level and that between LO and LS at the 5% level. The coefficient *rho* shows a significant correlation between LD and LV, and LD and LS only at the 10% level, while no significant correlation has been found for the pair LO-LV, though the *rho* value is close to the critical coefficient value at the 10% level. The Spearman correlation coefficient was also calculated for LV and the length of the text (i.e., the number of words in the text). LV and the length of the text are negatively correlated at the 5% significance level.

DISCUSSION

Correlational statistics like Spearman's *rho* may provide a statistically significant result without, however, any implication as to a cause-effect type of relation. If one wanted to find a causal type of relationship, it would be necessary to look for the contextual parameters such as the type of a speech event, conversation participants, setting, as well as the topics raised. This seems especially important in the case of spontaneous language of conversation, which seems particularly sensitive to all contextual changes. It may be both interesting and revealing to find out more about the speaker and the contexts of conversation with reference to the ten samples analysed above:

Group I (GI) -- 'good' (active and interesting) conversationalists

1) female academic, age 45; informal talk with a younger male colleague; topics: literature, friends

2) male social worker, age late 20s; informal gathering with friends (2 females, 1 male - the same age); the friends's flat; topics: holidays, entertainment, wine

3) female academic, age 55; informal gathering at the flat of another female academic (age 50) and her younger academic friend (female, age 25); topics: new decoration, art, painting

4) male legal civil servant, age 38; informal meeting with a friend (male architect, age 43); topics: maps, army, war and peace, human condition

5) male University administrator, age 55; semi-formal meeting with 5 male academics (age 40-55); topics: administrative worries of departments, library space, exams, etc.

Group II (GII) -- 'poor' (dull) conversationalists

6) female housewife (age 50), informal talk with a friend (female academic, age 50) and their husbands (age 50); lowest LD - many supportive moves, no initiating turns

7) male academic (age 48), informal meeting with a University colleague in his office; topics: students, curriculum, lectures

8) female prospective undergraduate, age 20; formal interview conducted by 2 male academics (age 40); topics: questions in English literature

9) female housewife (age 60), at her house; informal talk with a distant relative (female academic, age 40), visiting her; topic: long story of the children of a speaker's friend; (high verbosity, i.e., high LD and LV - closer to written narrative; low originality and creativity)

10) female secreary (age 21); informal chat with 3 other female secretaries (age 20); topics: gossiping about the boss, academics working in their departments and the courses they teach; (unexpectedly high LS explained by the fact that the speaker mentions some linguistic technical terms such as *sentence, nouns, verbs,* which have naturally low frequency counts in Thorndike and Lorge).

CONCLUSION

The results obtained in the present analysis cannot be treated as absolutely valid in all contexts. However, the statistics used help establish certain lexical relationships, potentially indicative of the tendencies occurring in the conversational English. Further statistical analysis aiming at capturing the nature of linguistic activeness and fluency must be based on larger and more numerous samples of the language. This further study may require some more sophisticated methods such as multiple correlation or cluster analysis.